HHS Public Access

Author manuscript

Sex Transm Dis. Author manuscript; available in PMC 2018 January 10.

Published in final edited form as:

Sex Transm Dis. 2017 August ; 44(8): 495–497. doi:10.1097/OLQ.0000000000000635.

An Illustration of Errors in Using the *P* Value to Indicate Clinical Significance or Epidemiological Importance of a Study Finding

Joseph Kang, PhD, **Jaeyoung Hong, PhD**, **Precious Esie, MPH**, **Kyle T. Bernstein, PhD**, and **Sevgi Aral, PhD**

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Atlanta, GA

Abstract

We conducted a simulation study to illustrate that *P* values can suggest but not confirm statistical significance; and they may not indicate epidemiological significance (importance). We recommend that researchers consider reporting effect sizes as *P* values in conjunction with confidence intervals or point estimates with standard errors to indicate precision (uncertainty).

Since 1999, experts have written about the inappropriate use of the *P* value to make judgments about the scientific significance (importance) of research findings in leading medical and scientific journals.[1–6] The primary concern is that a *P* value computed in a statistical significance test does not contain information about the clinical significance—the importance of an intervention—or epidemiological importance of the finding—a measure for the prevention and control of a disease in a population.[7,8] In 2016, the American Statistical Association issued a formal statement clarifying the proper use and interpretation of the *P* value and advising against using the *P* value to determine scientific significance of research findings.[9] The purpose of this article is to illustrate, with a simple simulated example, why small *P* values and narrow 95% confidence intervals do not indicate the clinical significance or epidemiological importance of a research finding. We recommend that authors report and interpret *P* values in conjunction with effect sizes and standard errors, or confidence intervals to support limited statements about the precision (uncertainty) and statistical significance of the findings. Conclusions about the clinical significance or epidemiological importance of research findings require clinical or epidemiological judgments that do not depend on statistical evidence alone.

METHODS

We devised an example of a prevalence difference known to be epidemiologically insignificant or unimportant. The hypothesis is to test the prevalence difference of 2 interventions—A and B. To compute the *P* value for the test, it is assumed that intervention A reduces the prevalence of an STD by 31% in group A and intervention B reduces the prevalence of the same STD by 27% in group B. Groups A and B are equal in size, and interventions A and B are equally effective. The relative effect of intervention A to that of

Correspondence: Joseph Kang, PhD, Mailstop E-02, Division of STD Prevention, Centers for Disease Control and Prevention, 12 Corporate Square Blvd, Atlanta, GA 30329. yma9@cdc.gov.

Conflict of interest: none declared.

intervention B is 1.21 as the prevalence odds ratio. We know from practical experience that a 4% difference in the effects of interventions A and B is not clinically significant or important. Let us suppose a 20% difference would be clinically important. Although 20% may be arbitrary, is comparable to the gender gap in 2014 gonorrhea rates—120.1 cases per 10^5 among men and 101.3 cases per 10^5 among women.[10] An intervention that closes that gap, that is, reduces the difference by 18.8%, would be epidemiologically important because closing the gap is a national goal. For our statistical simulation, the R statistical program[11] was used and is available as the supplementary document.

RESULTS

Using the outcomes described above, data can be readily simulated with different sample sizes. Figure 1 illustrates that as the sample size (*N*) and power increase, the *P* value becomes smaller, even when there is no change in the absolute difference of 4% (measure of effect). All the data points of this figure were generated with the same prevalence odds ratio of 1.21 and the absolute difference of 4%, as described in the previous section. For example, with a total sample size *N* of 100, $P = 0.24$ and appears not to be “statistically significant" using the standard threshold of $P < 0.05$. In contrast, a sample size *N* of 1800 results in $P = 0.002$, a value universally considered “statistically significant.” This phenomenon occurs because the *P* value is directly influenced by the sample size. As sample size increases, *P* values become smaller, crossing the 0.05 threshold to become “significant” regardless of whether the outcome is clinically significant. Without context, reporting only a *P* value for the group difference as evidence of its clinical significance or epidemiological importance will often result in misinterpretation. As stated above, the absolute difference in prevalence between the two groups in our example is 4%, a difference that is known to be unimportant. Thus, its effect size—the estimated difference of the prevalence rates (4%)—should be reported as well.

As shown in Figure 1, some small *P* values will show statistically significant differences. Instead of directly associating small *P* values and statistical significance with clinical significance, small *P* values should be interpreted as evidence supporting a rejection of the assumptions that the particular set of data are consistent with the proposed model for the data.[9] *P* values in Figure 1 were modeled under a null hypothesis assuming a prevalence ratio of 1.0. *P* values less than 0.05 mean the data are incompatible with the model’s null hypothesis assuming a prevalence ratio of 1.0, which should be rejected.[9] In the example provided, this is the only information the *P* value can provide. The incompatibility of the data and the statistical model’s null hypothesis indicated by *P* values less than 0.05 provides justification for a preliminary, not definitive or final, rejection of the null hypothesis.

Figure 2 shows that as the sample size *N* increases, the confidence interval for the effect size becomes narrower. With N =1800, confidence intervals for the estimated prevalence rates are reported as 31% to 37% and 24% to 30%, respectively. Equivalently, an estimated prevalence ratio of 1.38 can be reported with its confidence interval of 1.13 to 1.69 or its standard error of ±0.13. Neither statement about precision, however, justifies any statement about clinical significance or epidemiological importance.

DISCUSSION

In this article, we illustrated that *P* values measure statistical significance, but not necessarily clinical significance or epidemiological importance. The evaluation of public health interventions requires more careful epidemiological investigations including the assessment of the magnitude of attributable risks than simply reporting *P* values. When the *P* value was first proposed by Ron Fisher as an index of statistical significance, it was never meant to be used as an index of clinical significance or epidemiological importance.[12] The threshold of 0.05 was intended to serve as an initial or preliminary indicator of potential statistical significance, neither final nor confirmatory. Some disciplines, such as genomics,[13] have used *P* values to support some important discoveries, but the nature of the analysis in this respect is exploratory, rather than confirmatory. Even when the *P* value is meant to be exploratory, the use of 0.05 as a threshold may be misleading. Observe the horizontal dotted line at a significance level of 0.05 in Figure 1. If the dotted line was drawn at a significance level of 0.1, a larger number of simulated experiments would have statistically significant *P* values. Regardless of statistical significance levels, however, the prevalence difference (4%) or prevalence ratio (1.21) remains clinically insignificant and epidemiologically unimportant.

As shown in the simple example presented in this article, evidence of statistical significance is not evidence of clinical or epidemiological importance. Smaller *P* values are not necessarily associated with larger or more important effects, and larger *P* values are not necessarily associated with clinical insignificance or epidemiological importance of the effect. Recall that our example had a prevalence difference of 4%, but *P* values in Figure 1 varied. Thus, any analysis with a large sample size or high precision may produce a small *P* value, whereas analyses with small sample sizes or imprecise measurements may produce large *P* values even though the clinical or epidemiological effect may be important.

According to the ASA, “Cherry-picking effect sizes with small *P* values, also known by such terms as data dredging, significance chasing, significance questing, selective inference and “*P* hacking,” leads to spurious significant results.”[9] One way to prevent spurious findings is to report both effect sizes and corresponding confidence intervals so that readers can decide for themselves if the difference (or ratio) in effect is big enough to be meaningful or important on the basis of clinical or epidemiological criteria of significance. In some cases, however, effect sizes are not applicable to the result of the statistical test—e.g., a goodness of fit test only yields a *P* value. Therefore, there is still utility in using a *P* value to make decisions about a model and it should not be blindly abandoned.

Scientific writing tends to convey *P* values as statements about the truth of a null hypothesis, or about the probability that random chance produced the observed data. Yet, as shown in our study, whether *P* values are statistically rejected or not, the assumed clinical insignificance does not change. Evaluating the clinical significance of a finding is quite different from assessing the statistical significance.

Generally, *P* values can support exploratory or preliminary judgments about statistical significance, but they are neither certain, nor confirmatory, nor final. Without additional information, such as the effect size, a confidence interval, or standard error for context,

readers of scientific reports do not have sufficient information in the *P* value alone to make judgments about clinical significance or epidemiological importance of the statistical finding, as clearly illustrated in our example. Recent increases in big data analytics in health science provide substantially large data sets that can produce small *P* values that may not be epidemiologically useful. Indeed, the influx of modern-day big data in population health for different epidemiological disciplines would require more than the *P* value to assess the importance of new discoveries. To make study findings easier to interpret and to enable readers to make judgments about their usefulness, we recommend that in conjunction with *P* values, researchers also report confidence intervals or effect sizes with standard errors.

Acknowledgments

This project was supported in part by an appointment to the Research Participation Program at the National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention (CDC), administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and CDC.

References

1. Chavalarias D, Wallach J, Li A, et al. Evolution of reporting p values in the biomedical literature, 1990–2015. JAMA. 2016; 315:1141–1148. [PubMed: 26978209]
2. Baker, M. [Accessed October 5, 2016] Statisticians issue warning over misuse of P values. Available at: http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503
3. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999; 130:995–1004. [PubMed: 10383371]
4. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31:337–350. [PubMed: 27209009]
5. Greenland S, Poole C. Living with statistics in observational research. Epidemiology. 2013; 24:73–78. [PubMed: 23232613]
6. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych. 2015; 37:1–2.
7. Dixon P. The p-value fallacy and how to avoid it. Can J Exp Psychol. 2003; 57:189–202. [PubMed: 14596477]
8. Hunter JE. Development of green hospitals home and abroad. Psychol Sci. 1997; 8:3–7.
9. American Statistical Association. [Accessed October 5, 2016] Statment on statistical significance and p-values. Available at: http://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
10. Centers for Disease Control and Prevention. [Accessed January 23, 2017] Sexually Transmitted Diseases Surveillance: Gonorrhea. 2014. Available at: https://www.cdc.gov/std/stats14/gonorrhea.htm
11. R: A language and environment for statistical computing [computer program] 3.3.2. Vienna, Austria: R Foundation for Statistical Computing; 2016.
12. Nuzzo, R. [Accessed October 5, 2016] P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume. Available at: http://www.nature.com/news/scientific-method-statistical-errors-1.14700
13. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]

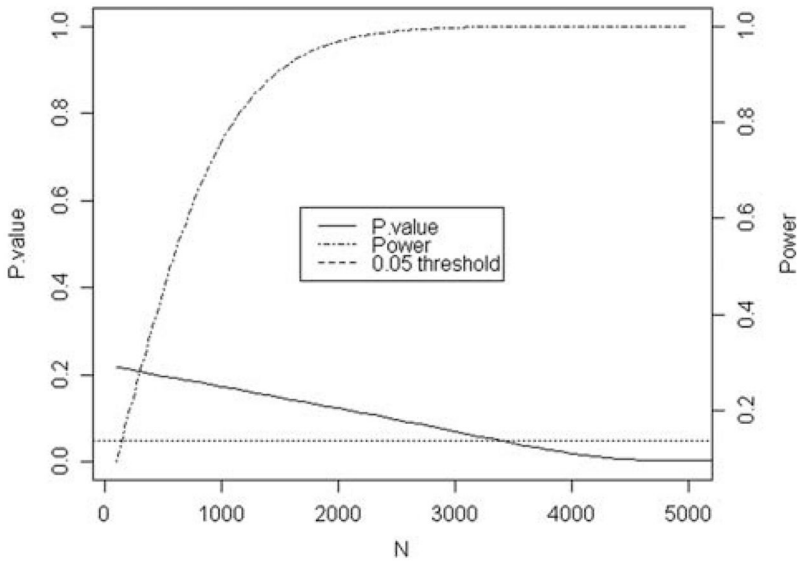

Figure 1.

P value and power versus sample size simulation with the same absolute difference of 4%.

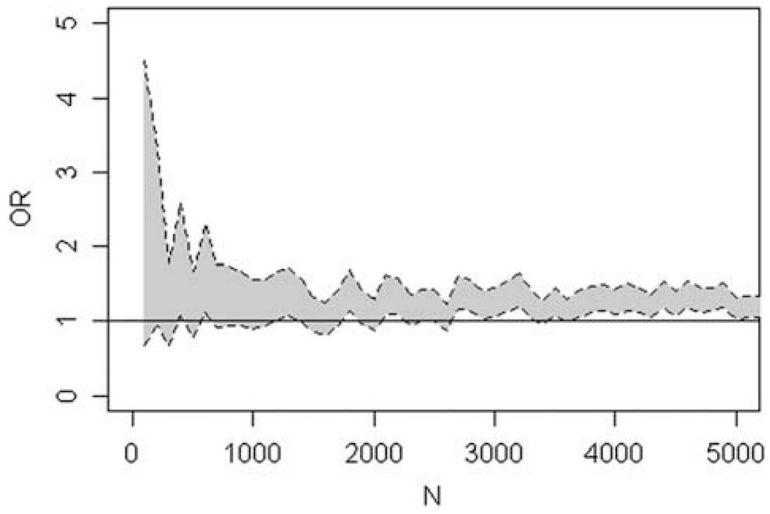

Figure 2.
Confidence interval for odds ratio (OR) versus sample size with the same absolute difference of 4%.